How to Embrace Your Spectrum for Faster Iterative Results*

John de Pillis
University of California
Riverside, California 92521

Submitted by Robert J. Plemmons

ABSTRACT

Given linear invertible $A: H \to H$ where $Ax = b$, write $A = A_0(I - B)$ where A_0^{-1} is easy to find. Define sequences $x_{n+1} = Bx_n + A_0^{-1}b$ and $y_{n+2} = (1+\lambda\mu^2)By_{n+1} - \lambda\mu^2 y_n + (1+\lambda\mu^2)A_0^{-1}b$, for arbitrary x_0, y_0, y_1 in H. Theorem 4.2 presents the algorithm for finding scalars λ and μ so that $y_n \to x$ faster than $x_n \to x$. In fact, if the spectrum of the iteration matrix B lies anywhere in the infinite vertical strip $\{z: |\mathrm{Re}(z)| < 1\}$, then $y_n \to x$ is assured. The algorithm follows:

(1) Capture the spectrum of B in a symmetric ellipse whose major semiaxes, M_r and M_i, lie in the real and imaginary axes, respectively.
(2) Define $\lambda = (M_r - M_i)/(M_r + M_i)$.
(3) Define μ to be the unique root in the $(0, 1)$ interval of the quadratic $(M_r + M_i)(1+\lambda\mu^2) = 2\mu$.

Then the asymptotic rate of convergence for $\{y_n\}$ is $R_y = -\log \mu$ [whereas the asymptotic rate for $\{x_n\}$ is well known to be $R_x = -\log \rho(B)$, where $\rho(B)$ is the spectral radius of B].

1. INTRODUCTION

Our objective is to find x in Hilbert space H where $Ax = b$ for given bounded linear operator (or matrix) A and fixed $x \in H$. We present a pattern for so-called one-part stationary iterative methods (or stationary methods of degree one); generalizing this pattern to two-part schemes, we will be able to run the race between the one-part and two-part methods and explore conditions under which the two-part methods provide faster convergence than, or accelerations of, the "standard" one-part schemes. The paper

*Research partially supported by AFOSR 78-2858D.

52 Vanderbilt Ave., New York, NY 10017 0024-3795/80/060125+19$01.75

concludes by showing how our two-part method relates to the nonstationary semiiterative methods of Golub and Varga [3] Varga [9], and Manteuffel [4], and to the "two-parameter" methods of Neithammer [6, 7].

Let us preview our main results. We begin by presenting a format for the "standard" one-part strategies which lends itself nicely to our developed format for two-part methods. We assume throughout, unless otherwise noted, that a certain easy-to-invert bounded linear operator A_0 has been selected. With this choice of A_0:

(1) Produce iteration operator or matrix B via the *multiplicative* splitting

$$A=A_0(I-B). \tag{1.1a}$$

(2) Obtain the operator A_1 via the *additive* splitting

$$A=A_0-A_1. \tag{1.1b}$$

(It follows that $A_1=A_0B$.)

(3) Use the additive splitting (1.1b) for arbitrary $x_0\in H$ to define the sequence $\{x_n\}$ by

$$A_0x_{n+1}-A_1x_n=b,$$

or equivalently

$$x_{n+1}=Bx_n+A_0^{-1}b, \qquad n=0,1,2,\ldots. \tag{1.1c}$$

(4) Use the multiplicative splitting (1.1a) to measure the asymptotic convergence rate R_x of the sequence $\{x_n\}$. In particular, if $\rho(B)$ is the spectral radius of B, then

$$R_x=-\log_{10}\rho(B). \tag{1.1d}$$

Roughly speaking, $1/R_x$ represents, asymptotically, the number of iterations in (1.1c) which produce one additional decinal place of accuracy in the x_n's.

With (1.1a) through (1.1d) as a template, we present our generalized scheme for stationary two-part splittings (or stationary second-degree methods) which, since they use the same invertible A_0 above, may be thought of as acceleration schemes for the one-part scheme (1.1a) through (1.1d). This is how we do it. With A_0 an easy-to-invert operator in hand:

(1) Produce (iteration) operators B_1 and B_2 via the *multiplicative* splitting

$$A=A_0(I-B_1)(I-B_2). \tag{1.2a}$$

(2) Obtain operators A_1 and A_2 via the *additive* splitting

$$A = A_0 - A_1 - A_2, \tag{1.2b}$$

where

$$A_1 = A_0(B_1 + B_2)$$

and

$$A_2 = -A_0(B_1 B_2).$$

(3) Use the additive splitting (1.2b) for arbitrary $y_0, y_{1 \in H}$ to define the sequence $\{y_n\}$ by

$$A_0 y_{n+2} - A_1 y_{n+1} - A_2 y_n = b,$$

or equivalently

$$y_{n+2} = (B_1 + B_2)y_{n+1} - (B_1 B_2)y_n + A_0^{-1} b \qquad \text{for all} \quad n = 0, 1, 2, \ldots \tag{1.2c}$$

(4) Use the multiplicative splitting (1.2a) to measure the asymptotic convergence rate R_y of sequence y_n. In particular,

$$R_y = \min\{-\log_{10}\rho(B_1), -\log_{10}\rho(B_2)\}. \tag{1.2d}$$

The scheme (1.2a)–(1.2d) is developed and justified for k-part splittings in [1].

Notice that once we have constructed sequences $\{x_n\}$ and $\{y_n\}$ according to (1.1c) and (1.2c), respectively, we can compare their convergence speeds by use of R_x and R_y as given by (1.1d) and (1.2d).

We now preview our main result:

THEOREM 4.2. *To find x such that $A_0(I - B)x = b$, define the two-part sequence $\{y_n\}$ for arbitrary initial y_0, y_1 by*

$$y_{n+2} = (1 + \lambda\mu^2)By_{n+1} - \lambda\mu^2 y_n + (1 + \lambda\mu^2)A_0^{-1} b \qquad \textit{for} \quad n = 0, 1, 2, \ldots, \tag{1.3}$$

where the scalars λ and μ are well defined by the configuration of $\sigma(B)$, the

spectrum of the iteration operator B (for which see the Abstract). Then $y_n \to x$ *whenever* $|\mathrm{Re}(\sigma(B))| < 1$ *with asymptotic convergence rate* $R_y = -\log_{10}\mu$.

Now if we have the one-part sequence on hand where, for initial arbitrary x_0, we have

$$x_{n+1} = Bx_n + A_0^{-1}b, \qquad n = 0, 1, 2, \ldots,$$

we proceed to ask the following questions and set forward their answers; to wit:

Q: Where must $\sigma(B)$ lie to insure that $y_n \to x$ where $Ax = b$?

A: If $\sigma(B)$ lies within the infinite vertical strip $\{z: -1 < \mathrm{Re}(z) < 1\}$, then $y_n \to x$. [By contrast, $\sigma(B)$ must lie within the unit circle in order that $x_n \to x$.]

Q: Is it difficult to find optimal λ and μ in (1.3)?

A: As described in the Abstract, we need to know the major and minor semiaxes (or an estimate) of a certain capturing ellipse for $\sigma(B)$ from which the scalars λ and μ are easily computed.

Q: If $\sigma(B)$ lies in the unit circle (so that $x_n \to x$), then does $y_n \to x$ faster?

A: In case the capturing ellipse for $\sigma(B)$ is a circle, then our two-part sequence (1.3) reduces to the two-part sequence (1.1c); this is the case $\lambda = 0$. But, whenever the embracing ellipse has a nonzero eccentricity ($\lambda \neq 0$), then $y_n \to x$ and its convergence rate is *always* faster than that of $x_n \to x$.

Q: How does (1.3) relate to the second-degree methods based on use of Chebyshev polynomials as seen in the works of Varga [9], Golub and Varga [3], Manteuffel [4], and Neithammer [5–7]?

A: It was Golub and Varga who showed that when A is positive definite (or $A_0 = I$ and B is hermitian convergent) then the Chebyshev semiiterative method, a nonstationary method of second degree, converges to (1.3) for the special case $\lambda = 1$; convergence here is to mean that in any machine with finite-length registers, constants of the Chebyshev nonstationary scheme will be indistinguishable from those of (1.3) (with $\lambda = 1$) after a certain time. In [9], Varga presents a result under hypotheses weaker than requiring that $A = I - B$ be positive definite; he requires only that $\sigma(B)$ be real and less than one in absolute value. By then applying SOR to a related system in a direct sum of vector spaces, (1.3) once again materializes (for $\lambda = 1$). To say that λ need not equal one in our scheme (1.3) is to allow $\sigma(B)$ to lie in the infinite vertical strip $\{z: |z| < 1\}$, and to guarantee convergence of $\{y_n\}$ to the solution x. Manteuffel considers those A whose spectrum lies in the right half plane [4]. Since A is continuous, this is tantamount to saying that $\sigma(A)$ lies in an infinite vertical strip $\{z: 0 < \sigma(A) < 2p\}$ for some scalar $p > 0$. By defining B by the equation $A = p(I - B)$ we see that $\sigma(B)$ lies in the infinite vertical

strip $\{z: -1 < \mathrm{Re}(z) < 1\}$ which satisfies the hypothesis of Theorem 4.2, rendering (1.3) a convergent sequence to the solution vector x. We will see in our final chapter how Manteuffel's analysis for the case where $\sigma(A)$ has one point can be generalized to the case where $\sigma(A)$ lies on a vertical line. Neithammer [6] studies two-part systems given by

$$y_{n+2} = (s_0 T + s_1) y_{n+1} + s_2 y_n + c$$

for certain scalars s_0, s_1, s_2. He presents elliptical regions for $\sigma(T)$ in order to insure convergence of sequence $\{y_n\}$; these turn out to be equivalent translations of our elliptical regions, although Neithammer's development is quite different in flavor from ours.

Q: If we know that the one-part sequence (1.1c) converges, is it practical or worthwhile to pass to a two-part sequence (1.2c) or (1.3)?

A: As for the *time* of execution, it turns out that the more eccentric or "flat" the embracing ellipse is for $\sigma(B)$, the more does the convergence rate R_y of (1.2d) exceed R_x of (1.1d). As for the *space* required, or extra memory, note that in (1.3) only one more vector need be stored than in the one-part algorithm of (1.1c). Significantly, if operators A_0 and B are already on hand, then no new *operators* need be computed for (1.3). Finally, if (1.1c) is developed toward the end of preserving sparsity, then (1.3) will also preserve sparsity.

The next two sections are devoted to developing a structure theorem for bounded operators on Banach space (Lemma 3.0) which seems to be of some independent interest. It says that any bounded operator B on Banach space is a scalar multiple of a sum of operators $U_1 + U_2$, where U_1 and U_2 both have spectral radii equal to one. Moreover, U_1 and U_2 may be chosen so that $U_1 U_2 = U_2 U_1 = aI$, a scalar operator. There is one technical constraint, namely that $\sigma(B)$ may not contain the foci of a certain capturing ellipse or else the spectral radii may have to deviate from unity. In finite dimensions, however, since $\sigma(B)$ has only finitely many points, we can get spectral radii as close to unity as we like.

We proceed now to the next two sections for a proof of this result.

2. FUNDAMENTAL LEMMA (VERSION I)

LEMMA 2.1. *Let $B: X \to X$ be a bounded linear operator on Banach space X. Then for any nonzero scalars a and λ where $4a^2\lambda$ is not in the spectrum of*

B^2, there exists invertible $V_{a,\lambda}: X \to X$ such that

$$B = a\left(V_{a,\lambda} + \lambda V_{a,\lambda}^{-1}\right). \tag{2.1}$$

Proof. It suffices to exhibit $V_{a,\lambda}$ as an analytic function of B. That is, we define a function $f_{a,\lambda}$ which is analytic on an open set $U \supset \sigma(B)$ such that for all $b \in \sigma(B)$ we have

$$b = a\left(f_{a,\lambda}(b) + \frac{\lambda}{f_{a,\lambda}(b)}\right). \tag{2.2}$$

With such an $f_{a,\lambda}$ in hand, we may use the operational calculus [2,8] to write

$$V_{a,\lambda} = f_{a,\lambda}(B) = \frac{1}{2\pi i}\int_C \frac{f_{a,\lambda}(z)}{z - B}\,dz. \tag{2.3}$$

It follows that $\sigma(V_{a,\lambda}) = f_{a,\lambda}(\sigma(B))$, or

$$\sigma(V_{a,\lambda}) = \{f_{a,\lambda}(b) : b \in \sigma(B)\}. \qquad \blacksquare \tag{2.4}$$

Can an analytic $f_{a,\lambda}$ be found which satisfies (2.2)? If we take the principal part of the square-root function, then the quadratic formula, in conjunction with (2.2), gives us a well-defined function

$$b \to f_{a,\lambda}(b) \qquad \text{for all} \quad b \in \sigma(B), \tag{2.5}$$

where

$$f_{a,\lambda}(b) = \frac{b + (b^2 - 4a^2\lambda)^{1/2}}{2a}.$$

Note that our hypothesis $4a^2\lambda \notin (B^2)$ assures us that the quadratic formula (2.5) is analytic on an open set containing $\sigma(B)$; the function $f_{a,\lambda}$ of (2.5) extends analytically so that operator $V_{a,\lambda}$ exists as per (2.3), whence (2.1) is satisfied.

REMARK. This form of the lemma, Version I, says that we may find scalars a and λ so that decomposition (2.1) obtains. The following section develops Version II, which says, in effect, that scalars a and λ may be

prechosen in accordance with the configuration of the spectrum of B so that certain spectral constraints on $V_{a,\lambda}$ ensue. (In particular, we want the spectral radii of $V_{a,\lambda}$ and of $\lambda V_{a,\lambda}^{-1}$ to be equal to one.)

3. FUNDAMENTAL LEMMA (VERSION II)

LEMMA 3.2. *Let $B: X \to X$ be bounded linear operator on the Banach space X. Then scalars $a > 0$, $\lambda \neq 0$ may be found along with bounded linear $V_{a,\lambda}: X \to X$ such that*

$$B = a\left(V_{a,\lambda} + \lambda V_{a,\lambda}^{-1}\right), \tag{3.1}$$

where

$$\rho\left(\lambda V_{a,\lambda}^{-1}\right) = \rho\left(V_{a,\lambda}\right) = 1, \tag{3.2}$$

may be found so that (3.1) *and* (3.2) *obtain if, moreover,* $4a^2\lambda \notin \sigma(B^2)$.

Proof. Lemma 3.1 tells us that for any prechosen $a > 0$, $\lambda \neq 0$, we may construct $V_{a,\lambda}$ so that at least (3.1) obtains. We now describe an algorithm for selecting $a > 0$ and $\lambda \neq 0$ in advance, based on the configuration of $\sigma(B)$, so that (3.2) holds true as well.

Recall that in (3.1), $V_{a,\lambda}$ is a certain analytic function of B, so that for all $b \in \sigma(B)$, there is some $v \in \sigma(V_{a,\lambda})$ such that

$$b = a\left(v + \frac{\lambda}{v}\right). \tag{3.3}$$

[This v is one of the two possible values of $f_{a,\lambda}(b)$ in (2.5).] Writing v in polar form, we have $v = |v|(\cos\theta + i\sin\theta)$. Substituting for v in (3.3), and writing complex $b = (\text{real}, \text{imaginary})$ in ordered-pair format, yields

$$b = \left(a\left[|v| + \frac{\lambda}{|v|}\right]\cos\theta, a\left[|v| - \frac{\lambda}{|v|}\right]\sin\theta\right). \tag{3.4}$$

Now as θ (which for the moment depends on b) is allowed to vary, (3.4) describes an ellipse $E_{|v|}(a, \lambda)$ in the complex plane with the following

properties:

$$\begin{aligned}&\text{The ellipse } E_{|v|}(a,\lambda) \text{ is centered at the origin.}\\&\text{The major semi-axis is } M_{|v|}(a,\lambda)=a[|v|+|\lambda/|v||].\\&\text{The minor semi-axis is } M_{|v|}(a,\lambda)=|a[|v|-|\lambda/v|]|.\\&\text{The foci are } 2a|\lambda|^{1/2} \text{ units from the origin, and lie on the real (imaginary) axis if } \lambda>0\ (\lambda<0).\end{aligned} \tag{3.5}$$

What (3.5) tells us, then, is that for any fixed $a>0$ and $\lambda\neq 0$, the spectrum $\sigma(B)$ is covered by a well-defined family of confocal ellipses. Moreover, each ellipse containing a $b\in\sigma(B)$ gives us information about $|v|$ through measurement of its major and minor semiaxes. [Solve for $|v|$, knowing the expression for measured $M_{|v|}(a,\lambda)$ or $M_{|v|}(a,\lambda)$.]

Here is the algorithm for pre-determining $a>0$ and real $\lambda\neq 0$.

$$\begin{aligned}&(1)\ \text{Construct a "capturing" ellipse for } \sigma(B) \text{ in accordance with (3.5); that is, an ellipse symmetric about both axes which contains points of } \sigma(B) \text{ in its interior. The foci should not intersect } \sigma(B).\\&(2)\ \text{With measurements of the major and minor semiaxes in hand } (M_1 \text{ and } m_1, \text{ respectively}), \text{ define } a>0 \text{ and } \lambda\neq 0 \text{ by } a=(M_1+m_1)/2,\ |\lambda|=(M_1-m_1)/(M_1+m_1), \text{ where } \lambda>0\ (\lambda<0) \text{ if the foci are on the real (imaginary) axis.}\end{aligned} \tag{3.6}$$

It remains only to verify that this choice of $a>0$ and $\lambda\neq 0$ constrains $|v|$ for all $v\in\sigma(V_{a,\lambda})$ by the inequality

$$|\lambda|\leqslant|v|\leqslant 1, \tag{3.7}$$

which will finally establish (3.2), and hence the lemma.

First, the largest (capturing) ellipse tells us that equality obtains in (3.7). This can be seen by solving for $|v|$ in

$$M_1=a\left(|v|+\left|\frac{\lambda}{v}\right|\right),$$

or

$$m_1=\pm a\left(|v|-\left|\frac{\lambda}{v}\right|\right)$$

[see (3.5)], using the particular values of a and λ in (3.6). The only eligible values for $|v|$ are

$$|v|=1 \quad \text{and} \quad |v|=|\lambda|. \tag{3.8}$$

Now all other (interior) ellipses have smaller major and minor semiaxes than does the capturing ellipse. From (3.5) we have $a[|v|+|\lambda/v|]<M_1$ and $\pm a[|v|-|\lambda/v|]<m_1$. These two inequalities imply the single inequality: For fixed $a>0$, real $\lambda\neq 0$,

$$\left||v|+\frac{\lambda}{|v|}\right|<1+\lambda. \tag{3.9}$$

Note, from (3.6), that $|\lambda|\leqslant 1$, so the right-hand side above is never negative. Squaring both sides of (3.9) and solving the resulting quadratic equation yields $|\lambda|^2<|v|^2<1$, or

$$|\lambda|<|v|<1,$$

which, in conjunction with (3.8), establishes (3.7), the lemma is proved. ■

4. THE PRINCIPAL THEOREM

We quote a special case of a recent result which will be needed in the proof of our main theorem.

THEOREM 4.1 [1, Theorem 5.1]. *Let $A: X\to X$ be a bounded linear invertible operator on a Banach space X. Consider the system $Ax=b$ where*

$$A=C_0(I-B_1)(I-B_2). \tag{4.1}$$

For arbitrary $y_0, y_1\in X$, define the two-part sequence $\{y_n\}$ by

$$y_{n+2}=(B_1+B_2)y_{n+1}-B_1B_2y_n+C_0^{-1}b, \qquad n=0,1,2,\ldots. \tag{4.2}$$

Then if $\rho(B_1)$ and $\rho(B_2)$, the spectral radii of B_1 and B_2, respectively, are

each less than one, we have

$$y_n \to x, \qquad \text{where} \quad Ax = b,$$

with asymptotic rate of convergence

$$R(\{y_n\}) = R_y = \min\{-\log B_1, -\log B_2\}. \tag{4.3}$$

REMARK. It is easy to check from (4.2) that if $\{y_n\}$ converges at all (say $y_n \to z$), then necessarily, $A_0(I - B_1 - B_2 + B_1B_2)z = b$. That is, $Az = b$, since $A = A_0(I - B_1)(I - B_2)$. Therefore, the limit of the sequence $\{y_n\}$ is the desired solution vector to our system.

Before proving our main result, let us recapitulate somewhat. Recall $A: X \to X$ is a bounded linear invertible operator on Banach X. We agree that A is of the form $A = A_0(I - B)$, A_0 is easy to invert.

Now let $\mathcal{E}$ be a capturing ellipse containing $\sigma(B)$, the spectrum of the iteration matrix B. Suppose $\mathcal{E}$ is symmetric about zero, and let the scalars M and m represent the major and minor semiaxes of $\mathcal{E}$, respectively. With this terminology in hand, we are ready to state our principal result.

THEOREM 4.2. *For linear bounded invertible $A: X \to X$ on a Banach space X, we have $A = A_0(I - B)$. Suppose we are to solve the linear system $Ax = b$. Then for arbitrary $y_0, y_1 \in X$, define the two-part sequence $\{y_n\}$ by*

$$y_{n+2} = (1 + \lambda\mu^2)By_{n+1} - \lambda\mu^2 y_n + (1 + \lambda\mu^2)A_0^{-1}b, \qquad n = 0, 1, \ldots, \tag{4.4}$$

where the scalars λ and μ derive from M and m, the major and minor semiaxes of the capturing ellipse $\mathcal{E}$, as follows:

$$\begin{aligned} \lambda &= \frac{M - m}{M + m} && \text{if the major semiaxis of } \mathcal{E} \text{ is horizontal,} \\ &= -\frac{M - m}{M + m} && \text{if the major semi-axis of } \mathcal{E} \text{ is vertical,} \end{aligned} \tag{4.5}$$

while μ is a solution to

$$\lambda\mu^2 - \frac{2\mu}{M + m} + 1 = 0. \tag{4.6}$$

If $\sigma(B)$, the spectrum of B, lies within the infinite vertical strip between $z=-1$ and $z=1$, i.e., if

$$\sigma(B)\subset\{z:-1<\operatorname{Re} z<1\}, \tag{4.7}$$

then for $\{y_n\}$ of (4.4), we have

$$y_n\to x, \qquad \text{where} \quad Ax=b,$$

with asymptotic rate of convergence

$$R(\{y_n\})=R_y=-\log\mu, \tag{4.8}$$

where μ defined by (4.6) is unique in the interval $(0,1)$.

Proof. Consider the operator

$$W=(I-\mu V)(I-\lambda\mu V^{-1}), \tag{4.9}$$

where the operator $V: X\to X$ and scalars μ and λ will be specified later. Equivalently, from (4.9) we have

$$W=(1+\lambda\mu^2)I-\mu(V+\lambda V^{-1}),$$

or

$$W=(1+\lambda\mu^2)I-\frac{\mu}{1+\lambda\mu^2}(V+\lambda V^{-1}). \tag{4.10}$$

Lemma 3.1 now allows us to specify the operator V above along with the scalars λ and μ. In fact, recall that $A: X\to X$ is of the form

$$A=A_0(I-B), \tag{4.11}$$

so in order to make W in (4.10) look like A in (4.11), we must ask under what conditions on V, λ, and μ we obtain the identity

$$B=\frac{\mu}{1+\lambda\mu^2}(V+\lambda V^{-1}). \tag{4.12}$$

It will suffice to establish (3.1) of Lemma 3.1. Thus, in order to render (4.12)

valid, we appeal to the algorithm of (3.6), which tells us to construct a symmetric capturing ellipse $\mathcal{E}$ with major and minor semiaxes equal to the respective values M and m. It then follows that λ is defined by (3.6) of the lemma, exactly in accordance with statement (4.5) of this theorem. Combining (3.1), (3.6), and (4.12), we obtain

$$a=\frac{\mu}{1+\lambda\mu^2}=\frac{M+m}{2}, \tag{4.13}$$

which is to say that μ is a solution to the quadratic equation

$$\lambda\mu^2-\frac{2\mu}{M+m}+1=0. \tag{4.14}$$

With scalars λ of (3.6) and a of (4.13) in hand, we may construct $V: X\to X$ according to (2.3), which establishes (2.1) [or its equivalent (3.1)], which, in turn, validates (3.1) with the extra guarantee from (3.2) that

$$\rho(\lambda V^{-1})=\rho(V)=1. \tag{4.15}$$

We are finally justified in relating A of (4.11) to W of (4.10) by substitution of (4.12) into (4.10). In fact, we obtain the string of equalities

$$\begin{aligned}
A&=A_0(I-B) && \text{from (4.11)}\\
&=A_0\frac{W}{1+\lambda\mu^2} && \text{from (4.10), (4.12)}\\
&=A_0\left(\frac{1}{1+\lambda\mu^2}\right)(I-\mu V)(I-\lambda\mu V^{-1}) && \text{from (4.9)}\\
&=C_0(I-B_1)(I-B_2), && (4.16)
\end{aligned}$$

where

$$C_0=A_0\left(\frac{1}{1+\lambda\mu^2}\right), \qquad B_1=\mu V, \quad \text{and} \quad B_2=\lambda\mu V^{-1}. \tag{4.17}$$

We may now appeal to our Theorem 4.1, since (4.16) establishes the requisite hypothesis (4.1). Then, by using the substitutions of (4.17), Theorem 4.1 assures us of the following: For arbitrary y_0, y_1 and X, the two-part

sequence y_n defined by

$$y_{n+2}=\mu(V+\lambda V^{-1})y_{n+1}-\lambda\mu^2 y_n+(1+\lambda\mu^2)A_0^{-1}b$$

$$=(1+\lambda\mu^2)By_{n+1}-\lambda\mu^2 y_n+(1+\lambda\mu^2)A_0^{-1}b \text{ [from (4.12)]},$$

has the property that $y_n\to x$, where $Ax=b$, with asymptotic rate of convergence

$$R(\{y_n\})=R_y=\min\{-\log(\mu V),-\log(\lambda\mu V^{-1})\} \quad \text{from (4.3)}$$

$$=-\log\mu \quad \text{from (4.15)}. \tag{4.18}$$

With y_{n+2} above subject to (3.6) and (4.13), our theorem is all but proved. We have only to show that the hypothesis (4.7) assures convergence of $y_n\to x$, where $Ax=b$. That is, we must only establish that (4.7), which constrains $\sigma(B)$ to the infinite vertical strip between $z=-1$ and $z=1$, assures us that there exists μ of (4.8) [see (4.18)] subject to (4.6) [see (4.14)] which is such that

$$0<\mu<1.$$

This follows from (4.6) [or from (4.14)] easily, once we define the quadratic g by

$$g(\mu)=\lambda\mu^2-\frac{2\mu}{M+m}+1$$

and observe that

$$g(0)=1>0, \tag{4.19}$$

while, from (3.6) we obtain

$$g(1)=\frac{2(M-1)}{M+m}<0 \qquad \text{if ellipse } \mathcal{E} \text{ is horizontal,}$$

$$=\frac{2(m-1)}{M+m}<0 \qquad \text{if ellipse } \mathcal{E} \text{ is vertical.}$$

We conclude from (4.9) that there exists a $\mu\in(0,1)$ such that $g(\mu)=0$.

Since $g'(\mu)<0$ for all $\mu\in(0,1)$, there exists only one $\mu\in(0,1)$ such that $g(\mu)=0$. ■

5. COMPARISON OF ONE-PART AND TWO-PART SEQUENCES

For convenience, let us restate the salient facts concerning our iterative sequences. With $A=A_0(I-B)$, we define the one-part sequence $\{x_n\}$ for arbitrary $x_0\in X$ by

$$x_{n+1}=Bx_n+A_0^{-1}b, \qquad n=0,1,2,\ldots. \tag{5.1}$$

Then $x_n\to x$, where $Ax=b$, if and only if $\sigma(B)\subset\{z:|z|<1\}$. The asymptotic convergence rate is

$$R_x=-\log\rho(B). \tag{5.2}$$

Define the two-part sequence $\{y_n\}$ for arbitrary $y_0, y_1\in X$ by

$$y_{n+2}=(1+\lambda\mu^2)By_{n+1}-\lambda\mu^2y_n+(1+\lambda\mu^2)A_0^{-1}b, \qquad n=0,1,2,\ldots. \tag{5.3}$$

Then $y_n\to x$, where $Ax=b$, if and only if $\sigma(B)\subset\{z:-1<\mathrm{Re}(z)<1\}$. If $\sigma(B)$ is contained in symmetric ellipse $\mathcal{E}$ with semiaxes M and m, then the asymptotic convergence is

$$R_y=-\log\mu, \tag{5.4}$$

where

$$\begin{aligned}\lambda&=\frac{M-m}{M+m} && \text{if } \mathcal{E} \text{ is horizontal,}\\ &=-\frac{M-m}{M+m} && \text{if } \mathcal{E} \text{ is vertical,}\end{aligned} \tag{5.5}$$

while $\mu\in(0,1)$ satisfies the equation

$$\lambda\mu^2-\frac{2\mu}{M+m}+1=0. \tag{5.6}$$

REMARK ($\lambda = 0$). Scalar λ behaves like the eccentricity of the capturing ellipse. That is, from the definition above,

$$-1 < \lambda < 1$$

and $\lambda = 0$ if and only if the eccentricity of $\mathcal{E}$, the capturing ellipse, equals zero. In fact, from (5.5) we see that $\mathcal{E}$ is a circle [$M = m = \rho(B)$] if and only if $\lambda = 0$, and in this case the two-part sequence (5.4) reduces to the one-part sequence (5.1). This is consistent with the fact that now $\mu = \rho(B)$ [from (5.6)], so that convergence rates R_x and R_y of (5.2) and (5.4) agree.

REMARK ($\lambda = \pm 1$). When $\lambda = 1$, then $\sigma(B)$ lies in the real interval $(-1, 1)$, whereas when $\lambda = -1$, $\sigma(B)$ is pure imaginary. In either case, when $|\lambda| = 1$, the capturing ellipse $\mathcal{E}$ is a degenerate ellipse, or line segment, having minor semiaxes m equal to zero [see (5.5)] and eccentricity equal to one. When $A = A_0(I - B)$ and $\sigma(B) \subset (-1, 1)$ (so that $\lambda = 1$), then our two-part sequence takes the form

$$y_{n+2} = (1 + \mu^2)\left[By_{n+1} - y_n + A_0^{-1} b \right] + y_n. \tag{5.7}$$

In Golub and Varga [3], we see that (5.7) is the limiting (stationary) equation of the Chebyshev semiiterative acceleration (nonstationary) equations. The proof of this result, involving the use of Chebyshev polynomials, also assumes that $A = A_0(I - B)$ is positive definite [hence, $\sigma(B) \subset (-1, 1)$]. Varga, in [6, p. 143], develops (5.7) by another route. There, it is assumed that $A = A_0(I - B)$, where $\sigma(B) \subset (-1, 1)$. Then by "imbedding" $Ax = b$ in direct-sum space and by applying SOR to the larger system, (5.7) is recaptured.

When $\lambda = -1$, then $\sigma(B)$ can be anywhere on the imaginary axis. For results in this situation see [1].

REMARK (Which is faster, $y_n \to x$ or $x_n \to x$). Assume that $\sigma(B)$ can be captured by (lies inside) a circle of radius $\rho(B) < 1$, so that *both* the one-part sequence (5.1) and the two-part sequence (5.3) converge. Then $\sigma(B)$ is also captured by a symmetric ellipse $\mathcal{E}$ having nonzero eccentricity. Then does $y_n \to x$ faster than $x_n \to x$ (is $R_y > R_x$)? This is easy to show when the capturing ellipse $\mathcal{E}$ lies wholly within the capturing circle. That is, if

$$m < M = \rho(B),$$

then consider the quadratic $g(\cdot)$ defined by

$$g(x) = \lambda x^2 - \frac{2}{M + m} x + 1.$$

Substitute for λ using (5.5), and it results that

$$\begin{aligned} &g(m)>0, \\ &g(\mu)=0 \qquad [\text{from (5.6)}], \\ &g(M)=g(\rho(B))<0. \end{aligned} \tag{5.8}$$

Since $g'<0$ on $(0,1)$, it follows that g is monotonically decreasing on $(0,1)$, which, in conjunction with (5.8), implies $\mu<\rho(B)$, i.e., $R_y>R_x$.

More precise comparisons of R_y with R_x when $\sigma(B)$ lies on an axis (so that the capturing ellipse lies wholly within the capturing circle) can be found in [1].

6. CONCLUDING REMARKS

We have been studying the equations

$$y_{n+2}=(1+\lambda\mu^2)By_{n+1}-\lambda\mu^2 y_n+A_0^{-1}b, \qquad n=0,1,2,\ldots. \tag{6.1}$$

Golub and Varga's Result

We have already noted that the Golub-Varga result [3] treats a nonstationary system which has as its limit the stationary system

$$y_{n+2}=(1+\omega)By_{n+1}-\omega y_n+A_0^{-1}b, \qquad n=0,1,2,\ldots, \tag{6.2}$$

where $0<\omega<1$, and $A_0(I-B)$ is positive semidefinite [see (5.7)]. One way to interpret our generalized scheme (6.1) is to say that (6.2) is the special case when $\lambda=1$ [which is to say that the embracing ellipse for $\sigma(B)$ is a real line segment] and $\omega=\mu^2$. But another way to view (6.1) is as a way of explaining why the Golub-Varga scheme (6.2) would converge with operators $A=A_0(I-B)$ which were *not* positive semidefinite. (We see such a phenomenon with the ADI method, where convergence theorems are established under one set of conditions, yet convergence seems to obtain under more general circumstances as empirical observation would indicate.) Although in (6.2) we treat ω as a scalar parameter deriving in some manner from positive semidefinite A, we see from (6.1) that $\omega=\lambda\mu^2$ represents a family of parameters [or a family of horizontal ellipses in the vertical strip $\{z:|\mathrm{Re}(z)|<1\}$]. In other

words, we now know the family of operators for which (6.2) will converge, namely those $A=A_0(I-B)$ where the spectrum of B lies inside the unit circle. Of course, ω will depend on the configuration of $\sigma(B)$.

Manteuffel's Result

Manteuffel's recent paper [4] gives us a nonstationary degree-two algorithm which uses Chebyshev polynomials in its analysis. The asymptotic convergence rate is specified and is obtained as a solution to a certain minimax problem. The basic assumption on the operator A which makes it more general than in the Golub-Varga result is that the spectrum of A lies in the right half plane; no assumptions about self-adjointness are needed.

How does this relate to our scheme (6.1)? From Theorem 4.2 we have a degree-two scheme which is stationary and whose asymptotic convergence rate is computed as a root of a certain quadratic (4.6). We further require in solving $Ax=b$ that $A=A_0(I-B)$ with $\sigma(B)$ a subset of the strip $\{z:|\mathrm{Re}(z)|<1\}$. But, on the other hand, to say that $\sigma(A)$ lies in the right half plane is equivalent to saying (since A is bounded) that for scalars $0<p<q$, $\sigma(A)$ is a subset of the strip $\{z: p<\mathrm{Re}(z)<q\}$. Once p and q are known, we may define operator B by writing $A=[(p+q)/2](I-B)$. It results, then, that $\sigma(B)$ lies in the infinite vertical strip between -1 and 1, so that such an A satisfies the hypotheses of our Theorem 4.2 and $Ax=b$ may be solved by use of (6.1). How does Manteuffel's minimax convergence rate compare with that given by (4.8) of Theorem 4.2? In [4, p. 323], we have the result, for example, that if $\sigma(A)$ consists of the single point $z_1=x_1+iy_1$, then the asymptotic convergence rate is

$$R_M=-\log\left(\frac{y_1}{x_1+\left(x_1^2+y_1^2\right)^{1/2}}\right).$$

Now set $p=q=x_1$, so that B is defined by $A=x_1(I-B)$. It follows that $\sigma(B)$ is the single point $-iy_1/x_1$. For the purposes of our analysis, using (6.1), we can embrace the single (imaginary) point $-iy_1/x_1$ with the same degenerate vertical ellipse which serves to embrace the entire vertical line segment $\{z=iy:|y|<|y_1/x_1|\}$. In other words, $\sigma(A)$ can be endowed with a vertical line spectrum as easily as a single-point spectrum, and the analysis and use of the two-part sequence (6.1) remain the same. What we have in either case is a vertical embracing ellipse for $\sigma(B)$ having zero (real) minor semiaxis and having (imaginary) major semiaxis equal to $|y_1/x_1|$. Consulting our algorithm (see Abstract or Theorem 4.2) we see that $\lambda=-1$, and so our convergence

rate is given through the logarithm of the root in the interval (0, 1), of the quadratic

$$\left|\frac{y_1}{x_1}\right|(1-\mu^2)=2\mu,$$

from which we also have that $-\log\mu=R_M$.

Neithammer's Results

In Neithammer's papers [5,6] (with different notation) he studies the two-part system (with some attention to k-part systems)

$$y_{n+2}=(s_0T+s_1)y_{n+1}+(s_2)y_n+(1-s_2)A_0^{-1}b, \qquad n=0,1,2,\ldots, \tag{6.3}$$

where T is a bounded linear operator and the s_i's are scalars subject to the constraints

$$s_0+s_1+s_2=1,$$

$$0<s_1<2,$$

$$|s_2|<1.$$

In answering the question where the spectrum of T should lie in order that (6.3) should converge, Neithammer [5] reveals that it suffices to have $\sigma(T)$ lie in an ellipse centered at $-s_1/s_0$, with

$$\text{real major semiaxis}=\frac{1-s_2}{s_0},$$

$$\text{complex major semiaxis}=\frac{1+s_2}{s_0}.$$

Another way to say this is that $\sigma(T)$ must lie in the infinite vertical strip between the real numbers $[-s_1-(1-s_2]/s_0$ and $[-s_1+(1-s_2)]/s_0$. Now compare (6.1) with (6.3) to obtain

$$s_0T=(1-s_2)B-s_1I,$$

which translates to saying that $\sigma(B)$ must lie in the infinite vertical strip

bounded between -1 and 1, exactly the hypothesis of our Theorem 4.2; i.e., (6.3) is an equivalent "translate" of (6.1).

In another work by Neithammer [7], $A=A_0(I-B)$ is assumed to have further properties (similar to being cyclic, having property A, or yielding knowledge of the optimal SOR relaxation factor). Moreover, $\sigma(B)$ is constrained to lie either in the real interval $(-1,1)$ or on the imaginary axis. Their is then generated a one-part sequence using SOR techniques where the resulting asymptotic convergence rate is about twice that of our two-part rate R_y of (5.4). This would demonstrate that having extra knowledge of the operator A or its structure can enhance convergence.

REFERENCES

1 J. de Pillis and M. Neumann, Iterative methods with k-part splittings, submitted for publication.

2 N. Dunford and J. T. Schwartz, *Linear Operators, Vol. I. General Theory*, Interscience, New York, 1958.

3 G. Golub and R. Varga, Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order Richardson iterative methods, Parts I, II, *Numer. Math.* 3:147–168 (1961).

4 T. A. Manteuffel, The Tschebychev iteration for nonsymmetric linear systems, *Numer. Math.* 28:307–327 (1977).

5 W. Neithammer, Iterationverfahren und allgemeine Euler-Verfahren, *Math. Z.* 102:288–317 (1967).

6 _____, Konvergenzbeschleunigung bei einstufigen Iterationsverfahren, *Internat. Ser. on Numer. Math* 15:235–240 (1970).

7 _____, On different splittings and the associated iteration methods, *SIAM J. Numer. Anal.* 16:186–200 (1979).

8 A. Taylor, *Introduction to Functional Analysis*, Wiley, New York, 1966.

9 R. Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1962.

Received January 1980; revised 27 April 1980